Applied Acoustics 142 (2018) 70–77

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition

Turgut Özseven

Department of Computer Engineering, Tokat Gaziosmanpaşa University, 60250 Tokat, Turkey

ARTICLE INFO

Article history:
Received 15 January 2018
Received in revised form 25 June 2018
Accepted 3 August 2018

Keywords:
Speech emotion recognition
Spectrogram
Texture analysis
SVM

ABSTRACT

Emotional state detection is an important part of human-machine interaction studies. The features used in emotion recognition are derived from the changes in facial mimics and speech signals. In emotion recognition from facial expressions, facial expressions are processed by image processing methods. If emotion recognition is performed via speech, speech is digitized by signal processing methods, and various features of speech are obtained by acoustic analysis. However, since the change in the features obtained by acoustic analysis is different according to emotion, the general success rate is changing. To overcome this limitation, the study of the effect of spectrogram images on emotional recognition is a current field of study. The purpose of this study is to investigate the effects of texture analysis methods and spectrogram images on speech emotion recognition. For this purpose, spectrogram images of speech were processed by four different texture analysis methods to obtain feature sets. The success rates for the emotion recognition of the obtained feature sets were experimentally investigated using support vector machines. In addition, the success of texture analysis methods was compared with acoustic analysis methods. The results have shown that texture analysis methods can be used for speech emotion recognition. When the results of the texture analysis methods were compared with those of the acoustic analysis, the texture analysis methods resulted in a 0.4% reduction in emotion recognition success rate. However, the combined use of both methods increased the success rate.

1. Introduction

Emotion is a physiological reaction that occurs in situations such as sadness, fear or happiness. Speech generation is also a physiological process. For this reason, the change in the emotional state will be reflected in the speech and face. This change in speech will also be an indicator of the person's identity, mental status and physical health.

One of the basic functions of emotion is the preparation of reactions that will be formed by emotion [1,2]. These reactions are softened with the help of a layer between the stimulus and the behavioural response [3,4]. Both the formation of emotion and the reaction are contributing to the learning of living things. Although these learnings are specific to the individual, the basic features are similar among all living things [5].

Speech Emotion Recognition (SER) keeps updating with being an old topic. First, in the mid-1980s, it was performed using the statistical properties of several acoustic parameters. In the following years, with the development of computer architecture, more complex emotion recognition algorithms have begun to be used [6]. For these evaluations, acoustic analysis is used for the processing of speech signals by digital signal processing methods and for objectively evaluating speech [7].

When the current studies are examined, acoustic analysis is used in the vast majority of SER studies. According to the results of the studies, the emotional speech data set, features and classifier preference cause a change in emotion recognition success. In addition, numerous studies using acoustic analysis, EMO-DB data and SVM classifier are included in the literature. In the study of the emotional states of the speakers speaking in front of the group, 276 acoustical and linguistic features were used and 84.84% success was achieved [8]. In the study where the Mel frequency cepstral coefficients feature set was used, 66% success was achieved with the SVM classifier [9]. In order to increase the success rates achieved with existing classifiers, researchers turned to hierarchical classifiers. In the study using SVM based 3DEC hierarchical classifier and 6552 feature set, 77.9% success was achieved [10]. A three-stage SVM classifier and MFCC feature set were used to achieve a 68.0% success rate [11].

E-mail address: turgut.ozseven@gop.edu.tr

https://doi.org/10.1016/j.apacoust.2018.08.003
0003-682X/

Researchers who think that the feature set can create adverse effects on the SER are aiming to increase SER performance while reducing the feature size using feature reduction methods. Chiou and Chen (2013) reduced the feature size above 6000 by 37, while maintaining 80% success [12].

As the emotional tracks of speech are reflected in the voice, it is expected to be reflected in the spectrogram of speaking. The results of the study using spectrogram images and deep convolutional neural network showed that deep learning methods and spectrogram images could be used successfully in SER and 73.8% of emotion recognition success was obtained [13]. In another study using Mel spectrogram and AlexNet deep learning model, features were collected by discriminant temporal pyramid mapping method. According to the results obtained, the pre-trained deep learning model performed well in emotional speech. When fine-tuned was not used and used, 72.35% and 82.65% success was achieved on EMO-DB respectively [14]. In the study using prosodic and spectral feature sets, 2185 features were used for seven emotions on EMO-DB, including 960 spectral, 241 prosodic, 780 harmonic energy and 204 spectrograms. The classification was made with a hierarchical classifier, which is broken down by gender and emotion. As a result, the success rate of men and women was 86.9% [15]. In the study, which presented a new algorithm for emotion recognition through spectrogram, artificial neural networks and five emotions (angry, sad, happy, neutral and fear) in EMO-DB were used. An average of 80.5% success was achieved for 5 emotions [16]. In another study on emotion recognition through spectrogram, spectrogram images were obtained for four emotions (anger, sadness, happiness and fear) in EMO-DB, eNTERFACE and KHUSC-EmoDB databases. A cubic curve is used to enhance the contrast of the spectrogram images. Later, a feature set was created using Laws masks to characterize the emotional state. The obtained feature set is classified with SVM classifier. The average success rate was 77.42% on EMO-DB, 73.06% on eNTERFACE and 65.20% on KHUSC-EmoDB [17].

Existing studies show the usability of spectrogram images for SER. However, in the studies, a limited number of texture analysis methods were used with different classifiers. This creates the question of how other methods will have an effect on the SER. In addition, the question of which method (spectrogram images or acoustic analysis) is more effective for the SER has not yet found an answer. The purpose of this study is to contribute the literature to answer these two questions. For this purpose, the success of different texture analysis methods on SER was investigated. The results obtained were compared with those obtained by acoustic analysis.

The study was organized as follows; Section 2 contains the used emotional speech database, Sections 3 and 4 contains the used methods. Section 5 contains the results of the study. The results obtained in Section 6 of the work are interpreted.

2. Emotional speech database

Three different datasets (EMO-DB, eNTERFACE05 and SAVEE) were used to improve the generalization ability of the findings obtained in the study. The Berlin Database of Emotional Speech (EMO-DB) is derived from the expression of different emotions by actors (anger, boredom, disgust, anxiety/fear, happiness, sadness, neutral). Audio recordings have 16 kHz sampling frequency and are 16 bit mono [18]. The eNTERFACE'05 is an audio-visual emotion database (anger, disgust, fear, happiness, sadness, surprise). The database contains 42 subjects, coming from 14 different nationalities [19]. Surrey Audio-Visual Expressed Emotion (SAVEE) Database recorded an audio-visual emotional database (anger, disgust, fear, happiness, sadness, neutral, surprise) from four native English male speakers, one of them was postgraduate student and rest were researchers at the University of Surrey [20]. The distribution of the data used in the study is given in Table 1.

3. Feature extraction

The spectrogram is a visual representation of the spectrum of signal frequencies that vary with time. That is, the data in the spectrogram relates to the frequency distribution of the speech signal. Texture analysis is concerned with defining the characteristics of structural features of images. The texture is defined as the spatial variation of pixel intensities. Texture analysis methods provide unique information about the spatial variation of texture or pixels. In this context, texture analysis methods are used in many areas ranging from face recognition to medical image processing [21]. In order to investigate the effectiveness of these methods on spectrogram images, Gabor Filter (GF), Histogram of Oriented Gradients (HOG), Gray Level Co-Occurrence Matrix (GLCM) and Wavelet Decomposition (WD) methods are used. Since the spectrogram is related to the frequency distribution, the acoustic properties to be used are determined by considering the frequency distribution and the fundamental frequency, formant frequencies and Mel-Frequency Cepstral Coefficient (MFCC) are used. The flow diagram for the feature extraction is given in Fig. 1.

3.1. Spectrogram based features

The intensity of the research on the image processing and the developed technologies have increased the calculation power of the image processing methods. To investigate the effects of these developments on the spectrogram images, the spectrogram image of each speech record was obtained with MATLAB. The sampling rate for speech records is 16 kHz. When spectrogram images were obtained, sampling frequency, windowing method, maximum amplitude value and overlap rate were used 16 kHz, hamming, 50 dB and 50%, respectively. In addition, spectrogram images have been converted to a graycolor tone to reduce workload. The signal and spectrogram image of the sample speech recordings are given in Fig. 2.

GF, HOG, GLCM and WD methods are used to the feature extraction from the spectrogram images. The coding of these methods is done with MATLAB.

GF is based on multi-channel filtering, which mimics some features of the human visual system [22]. GF consists of a series of Gaussian filters that cover the frequency domain in the spatial and frequency domain with different radial frequencies and orientations [22,23]. The Gabor core can be adjusted to the desired angle and wavelength so that the desired feature can be found on the image.

The 2D Gabor function $g(x, y)$ with respect to the input image $I(x, y)$ is transformed according to Eq. (1) and the resulting Gabor feature image $r(x, y)$ [23,24]:

$$r(x,y) = \iint_{\Omega} I(\xi,\eta)g(x-\xi,y-\eta)d\xi\eta \tag{1}$$

Gabor functions used for the convolution are given in Eqs. (2) and (3).

$$g_{\lambda,\theta,\varphi,\sigma,\gamma}(x,y) = exp\left(-\frac{x'^2+\gamma y'^2}{2\sigma^2}\right).\cos(2.\pi.\frac{x'}{\lambda}+\varphi) \tag{2}$$

$$x' = x.cos\theta + y.sin\theta y' = -x.sin\theta + y.cos\theta \tag{3}$$

In the formula; g is gabor core, σ is the standard deviation of the size of the region to be included in the weighted sum, λ wavelength, θ angle, φ phase angle and γ aspect ratio.